also not optimal because it does not include CD4+ cell count and viral load. Thus, we conclude that BASS has some, but not all, of the right pieces. Future research needs to be more focused with regard to the population under study and include measures of clinical judgment, and immunologic and virologic markers, as appropriate, if staging systems are to be clinically useful. □

Wafaa El-Sadr
Harlem Hospital Center/Columbia University
College of Physicians and Surgeons
New York, NY

James D. Neaton
School of Public Health
University of Minnesota
Minneapolis

References

1. Seage GR III, Gatsonis C, Weissman JS, et al. The Boston AIDS Survival Score (BASS): a multidimensional AIDS severity adjustment. *Am J Public Health.* 1997;87:567–573.
2. Justice AC, Feinstein AR, Wells CK. A new prognostic staging system for the acquired immunodeficiency syndrome. *N Engl J Med.* 1989;320:1388–1393.
3. Turner BJ, Markson LE, McKee L, et al. The AIDS-defining diagnosis and subsequent complications: a survival-based severity index. *J Acquir Immune Defic Syndr.* 1991;4:1059–1071.
4. Neaton JD, Wentworth DN, Rhame F, et al. Considerations in choice of a clinical endpoint for AIDS clinical trials. *Stat Med.* 1994;13:2107–2125.
5. Mocroft AJ, Johnson MA, Sabin CA, et al. Staging system for clinical AIDS patients. *Lancet.* 1995;346:12–17.
6. Justice AC, Aiken LH, Smith HL, Turner BJ. The role of functional status in predicting inpatient mortality with AIDS: a comparison with current predictors. *J Clin Epidemiol.* 1996;49:193–201.
7. Saravolatz L, Neaton JD, Sacks L, et al. CD4+ lymphocyte counts and patterns of mortality among patients infected with human immunodeficiency virus who were enrolled in the Community Programs for Clinical Research on AIDS. *Clin Infect Dis.* 1996;22:513–520.
8. Coombs RW, Welles SI, Hooper C, et al. Association of plasma human immunodeficiency virus type-1 RNA level with risk of clinical progression in patients with advanced infection. *J Infect Dis.* In press.
9. Welles SI, Jackson JB, Yen-Lieverman B, et al. Prognostic value of plasma HIV-1 RNA levels in patients with advanced HIV-1 disease and with little or no prior zidovudine therapy. *J Infect Dis.* In press.
10. Galetto-Lacour A, Yerly S, Perneger TV, et al. Prognostic value of viremia in patients with long-standing human immunodeficiency virus infection. *J Infect Dis.* 1996; 173:1388–1393.
11. Mellors JW, Rinaldo CR Jr, Gupta P, et al. Prognosis of HIV-1 infection predicted by the quantity of virus in plasma. *Science.* 1996;272:1167–1170.
12. Knaus WA, Harrell FE Jr, Lynn J, et al. The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. *Ann Intern Med.* 1995; 122:191–203.
13. Abrams KI, Goldman AI, Launer C, et al. A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. *N Engl J Med.* 1994;330: 657–662.

Annotation: What Can Be Done about Missing Data? Approaches to Imputation

Missing observations are a nuisance commonly encountered in research on human populations. Over the years, statisticians have developed a body of theory and methods for handling missing data. A key reference is the book by Little and Rubin.[1]

One approach to analyzing incomplete data is to fill in each missing item with an imputed value and analyze the data set as if it were complete. Although such methods can give unbiased estimates, standard errors are generally too small because they do not reflect uncertainty about the values of the missing observations. To address this problem, Rubin[2] proposed a procedure called *multiple imputation:* Using a predictive model, one simulates several (often five) completed versions of the data, analyzes each filled-in data set separately, and combines the analyses into a single summary analysis using simple averaging formulas. If the model is correct, confidence intervals and tests will properly reflect uncertainty from both the sampling and the incompleteness.

For example, suppose that a continuous confounder, call it X, is missing for some fraction of the subjects in a study whose goal is to estimate the effect of a factor Y on an outcome. If there were no missing data, one would execute a logistic regression predicting outcome from Y, X, and possibly other variables. If we exclude X from the logistic regression, we can use all the subjects, at the risk of some bias in the odds ratio for Y. If instead we exclude all cases with missing X values, we can use all the variables but not all the subjects; this strategy risks both bias and loss of power.

A multiple imputation analysis seeks to avoid these pitfalls by substituting model-based imputations for the missing X data. The analysis proceeds something like this: First, using only those subjects who have no missing data, estimate a multiple linear regression predicting X from other relevant variables. Second, combine these regression coefficients with the observed values of the predictors of X to compute predicted values for the subjects whose X values are missing. Third, use a random number generator to simulate a set of residuals, and add the residuals to the regression predictions to produce a single set of imputed X values. Repeating this process five times—each time simulating new residuals—yields a set of five imputations. To analyze the data, estimate the logistic regression coefficients and their standard errors separately for each imputed data set; this yields a set of five log odds ratios for Y, together with their standard errors. Using Rubin's formulas, combine the five estimates and standard errors into an overall estimate and standard error. If the fraction of missing data is small, the estimates and standard errors will be nearly the same across imputations, and the overall estimate and standard error will also be about the same. If the fraction of missing data is large, the five estimates may differ substantially, in which case the overall standard error can be much bigger than the individual standard errors.

Rubin originally proposed multiple imputation to handle missing data in sample surveys. The survey statisticians, working in concert with subject-matter experts, would build an imputation model using all available design information, possibly including confidential data. Public-use distributions of the data would consist of the observed items together with multiple sets of imputations for the missing items. Consumers of the data would obtain proper inferences by simply

Editor's Note. See related article by Gomel et al. (p 673) in this issue.

analyzing each imputed data set separately and combining the answers using Rubin's formulas.

Many researchers now use multiple imputation as an analysis method in its own right, without any intention of producing public-use data. An example can be seen in the paper in this issue by Gomel and colleagues.[3] As we have seen, multiple imputation involves simulating the missing observations under a predictive model. Not surprisingly, faulty model assumptions can lead to invalid results. Before commenting further on the paper by Gomel et al., I will consider some features of models for incomplete data.

The first component is the regression for predicting X from the fully observed variables. One can assess the adequacy of this model with standard diagnostic procedures, such as residual plots and goodness-of-fit tests. One should also inspect the imputed data; if predictions are absurd, such as systolic blood pressures greater than 200 for otherwise healthy subjects, we know that the model is not reliable.

A second component of the model is the *missing-data mechanism,* a description of the probability distribution of the pattern of missing observations. Often, one can think of this as a logistic regression model specifying the probability that an item is missing as a function of the data values. If the probability that an item is missing does not depend on the observed or missing values—imagine a research assistant shuffling the data forms and throwing away the bottom third—we say that the missing data are *missing completely at random* (MCAR). When MCAR holds, regressions using all complete records, means of available cases, nonparametric tests, and moment-based methods such as generalized estimating equations (GEE) are all valid.

If the probability that an observation is missing can depend on the values of observed items but not on the value of the missing item itself, we say that the missing data are *missing at random* (MAR). When MAR holds, maximum-likelihood estimates of model parameters are valid even if one does not simultaneously estimate the parameters of the missing-data mechanism. Many multiple imputation procedures implicitly assume MAR missingness.

If the probability that an item is missing depends on the unobserved value of the missing item itself, the missing-data mechanism is said to be *nonignorable.* To get correct analyses from such data, one must estimate the dependence of the missingness probability on the missing values.[4,5] Such estimates can be unreliable and difficult to compute, and are not yet available in statistical analysis packages.

Consider the measurement of body mass index (BMI) in Gomel et al. If body mass index data were missing because subjects happened to be out on calls at the time of scheduled measurements, the mechanism would be MCAR, and complete-case and available-case means and GEE regressions would be valid. Suppose, on the other hand, that body mass index is missing because subjects who had high body mass index values at earlier visits avoided being measured at later visits out of embarrassment, regardless of whether they had gained or lost weight in the intervening period. Such missing data would be MAR but not MCAR, and maximum-likelihood analyses such as the random effects models in SAS Proc Mixed (SAS Institute, Cary, NC) would be valid, whereas GEE would not. Suppose now that subjects were more likely to avoid being measured if they had put on extra weight since the last visit. Such data are nonignorably missing, and a correct analysis would require complicated modeling, as in Diggle and Kenward.[4]

Usually, the data will contain little information to help us decide whether the missing-data mechanism is MCAR, MAR, or nonignorable. Investigators should make every effort to determine why some observations are missing while others are not. Careful attention to this issue during data collection can help investigators make more informed model choices during the analysis phase.

A final point to consider is the amount and pattern of missing data. If only a few values are missing, then even single imputation can give reasonable results. Often, missing items are concentrated in a small number of variables, with scattered missing data on other variables. When this occurs, one can provisionally fill in the scattered missing data and concentrate modeling efforts on the variables where the fraction missing is greatest. In practice, the fraction of missing data that can be considered negligible depends on the intended analysis, the missing-data mechanism, and the pattern of missing data; consequently, there are no firm rules for deciding when it is worthwhile to undertake multiple imputation. Nevertheless, if the fraction of cases with missing observations is less than, say, 5%, and the mechanism is ignorable, most simple analyses should be satisfactory. If there is any doubt, one can execute a simple sensitivity analysis by producing two extreme imputations—one consisting of high values and the other of low values—and comparing their complete-data analyses to reveal the range of potential sensitivity. If this range is narrow, simple approaches should be satisfactory.

Multiple imputation has several practical advantages over other methods. First, it is easy to adapt multiple imputation to adjust for suspected nonignorability. In the body mass index example, one could add a constant, say 2 units, to the imputations from an ignorable model to reflect the fact that missing values may be systematically higher than observed values. If nonignorable and ignorable imputations give similar results, the ignorable model gains credibility. Second, if the prediction model is suspect, one can use an *implicit* imputation model as opposed to explicit models like that of Gomel et al. One class of implicit models is based on predictive-mean matching,[6] a technique that locates, for each subject with missing data, a set of subjects whose data are complete and that give similar predictions for the missing variable. Imputed values are selected randomly from the observed values of the complete-data matches, eliminating the possibility of nonsensical imputations. Finally, Rubin's formulas yield simple estimates of the fraction of information missing for each analysis. If the fraction is small, results are insensitive to the imputation model and method. If the fraction is large, many imputations are needed to obtain precise estimates, and results are perilously sensitive to model misspecification.

It is encouraging to see authors such as Gomel et al. making serious efforts to execute principled analyses of incomplete data. However, their terse presentation does not indicate how extensive the missing data are and how they handled ignorability and model fit. I recommend that investigators explicitly tabulate the quantities and patterns of missing observations. They should also give serious attention to questions of ignorability: Why are these observations missing, how are they likely to affect analyses, and is there a need for nonignorable imputation? And finally, investigators should consider whether their imputation models give sensible predictions. If they do not, the more robust matching techniques are preferable. □

Daniel F. Heitjan
School of Public Health
Columbia University
New York, NY

References

1. Little RJA, Rubin DB. *Statistical Analysis with Missing Data.* New York, NY: John Wiley and Sons, Inc; 1987.
2. Rubin DB. *Multiple Imputation for Nonresponse in Surveys.* New York, NY: John Wiley and Sons, Inc; 1987.
3. Gomel MK, Oldenburg B, Simpson JM, Chilvers M, Owen N. Composite cardiovascular risk outcomes of a work-site intervention trial. *Am J Public Health.* 1997;87:673–676.
4. Diggle P, Kenward MG. Informative dropout in longitudinal data analysis. *Appl Stat.* 1994;43:49–94.
5. Ibrahim JG, Lipsitz SR. Parameter estimation from incomplete data in binomial regression when the missing data mechanism is nonignorable. *Biometrics.* 1996;52: 1071–1078.
6. Heitjan DF, Landis JR. Assessing secular trends in blood pressure: a multiple-imputation approach. *J Am Stat Assoc.* 1994;89:750–759.

Topics for Our Times: Norplant Coercion—An Overstated Threat

The pervasive shroud of controversy surrounding the contraceptive Norplant is troubling. The safety and efficacy of the method were documented in research that led to approval by the Food and Drug Administration (FDA) in 1990 and were recently reaffirmed in statements by the International Medical Advisory Panel of the International Planned Parenthood Foundation[1] and the FDA.[2] Moreover, women using Norplant report high levels of satisfaction with the method.[3,4]

At the root of the controversy is the labeling of Norplant as an instrument of coercion. Within 2 days of the FDA approval, an editorial in the *Philadelphia Inquirer* suggested that Norplant should be used as a "tool in the fight against black poverty."[5] The effect was immediate: Norplant was cast as a method of social control. As Ellen Goodman, the syndicated columnist, keenly observed, "It took 24 years to develop, test, and approve an implantable device. . . . It took less than two weeks for Norplant to be billed as a new method of coercion."[6]

Fears about the coercive use of Norplant initially focused on public actions targeted at poor and minority women. Legislators in 13 states have proposed nearly two dozen bills designed to use Norplant as an instrument of social engineering—conditioning welfare payments on Norplant use, or enticing women on welfare to use Norplant through financial incentives.[7,8] Moreover, at least four women convicted of child abuse have had Norplant inserted as a condition of probation.[7,9] These actions are particularly insidious because they single out one class of women—poor, single mothers, who are frequently women of color—as the targets of fertility control. The fact that these policymakers specified Norplant as the agent of control lent credence to the charge that Norplant could be used as an instrument of coercion.

The contentiousness of these issues has overshadowed the fact that the reality of public coercion has been far less dramatic than its threat. No state legislature has enacted into law any of the proposals linking Norplant to welfare payments. Moreover, the forced use of Norplant in the judicial system seems to be abating. It has been made illegal in Illinois, and in California a judge is facing formal disciplinary charges for making Norplant insertion a requirement of probation.

These public attempts to force Norplant use raised suspicions of coercion in the private interactions between women and their health care providers. This would be the case if clinicians were violating principles of informed consent and pressuring women to use the method.[10–13] These concerns were particularly charged because of the involuntary sterilization of disabled, poor, and minority women as recently as the 1970s.[14]

The absence of data has made it impossible to empirically assess the claims of provider coercion. We can make such an assessment with data from our study of Norplant choice in the United States. This sample consists of over 2000 low-income women interviewed at the time that they were choosing a new contraceptive method. By design, we oversampled women choosing Norplant, who constituted 45% of the study sample. The remaining women chose Depo-Provera, oral contraceptive, or sterilization. Respondents were recruited from large, hospital-based, family planning clinics in Dallas, Pittsburgh, and New York City.

Our findings on coercion are clear. Women did not perceive that they were coerced in their choices regarding Norplant. When asked if they had felt any pressure from a health care provider to use Norplant, only three women in our sample of 2000 said yes. One of these chose sterilization—which suggests the reason for the pressure was to encourage her to avoid a permanent method of contraception. The other two women reported that they had to return to the clinic a number of times to obtain Norplant—which indicates that pressure was not associated with a rush to insert the method. The absence of coercion is further reflected in women's responses to a question probing why they chose Norplant. Only four women cited health-care-provider influence as a reason for their choice. Overwhelmingly, women pointed to the positive characteristics of Norplant—convenience, effectiveness, and duration—as the primary determinants of their choice.

Finally, the data show that the process of obtaining Norplant runs counter to the claim of coercion. A logical way to coerce Norplant use is to speed up the process of adoption so that women do not have time to change their minds. A comparison of how women obtained Norplant vs the pill indicates that this did not occur. First, Norplant adopters had to make significantly more visits to the clinic to obtain their method than did women seeking the pill. Second, those choosing Norplant were not rushed through counseling. They received an average of 42 minutes of counseling, 20 minutes more than women choosing the pill. Third, women rated the process of obtaining Norplant as significantly more difficult than that for the pill. All of these findings point to the fact that providers are not coercing women to initiate Norplant use in the large urban health care systems where most poor and minority women receive their family planning services.

The public debate about Norplant has been consumed by a focus on coercion. This outcry has functioned as a double-edged sword. It might well have served to reduce the magnitude of the problem by creating an atmosphere of vigilance. However, concerns regarding coercion have been overstated and have served to stigmatize the method. This has made it difficult for women and providers alike to impartially evaluate Norplant, a contraceptive option that we believe has the potential to serve many women well.

We do not view the controversy surrounding Norplant as an anomaly, but